MATH1712 Probability and Statistics II
Homework 5

`http://www1.maths.leeds.ac.uk/~voss/2017/MATH1712/`
Jochen Voss, J.Voss@leeds.ac.uk
2017/18, semester 2

This exercise sheet will be discussed in the tutorials of the week beginning 26th February.

The main topic of this exercise sheet is linear regression.

During the tutorial.

(1) Work together to find an estimator $\hat\beta$ for the parameter β in the simplified regression model $Y_i = \beta x_i + \varepsilon_i$ (with no intercept), using the least squares method.

(2) In lectures, we used a regression line of the form $y = \alpha + \beta x$ to predict y from x. Instead, we could have used $x = \gamma + \delta y$ to predict x from y. Is this the same model in a different form? Or a different model? How do we know which of the approaches to use?

(3) In the model $Y_i = \alpha + \beta x_i + \varepsilon_i$, the residuals ε_i can be seen as an error or uncertainty in the observations y_i; in contrast, we assumed the values x_i to be known exactly. How would you construct a model where there are uncertainties about the values of both x and y?

Homework questions. Your solutions to these questions contribute towards your final mark for the module. Please hand in your solutions **to your tutor** via the silver pigeon holes (down the stairs from the maths reception) by **Tuesday, 6th March, 5pm**.

Exercise 17. Assume that we have observed paired data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, and we want to describe these data using the model $Y_i = \gamma x_i^2 + \varepsilon_i$ where ε_i are i.i.d. with $\mathbb{E}(\varepsilon_i) = 0$.

a) Using the least squares method, determine an estimator $\hat\gamma$ for the parameter γ.

b) Show that your estimator $\hat\gamma$ is unbiased.

Exercise 18. Consider the following data set:

i	1	2	3	4	5	6	7	8	9	10
x_i	1.027	1.755	2.135	2.513	3.577	4.520	5.142	7.156	7.460	8.396
y_i	8.308	6.154	8.233	7.814	7.177	6.085	6.323	5.037	5.432	4.585

Our aim is to fit a linear regression model to this data.

a) Using the `lm()` function in R, estimate the intercept α and the slope β. (You can either type in the data manually, or download it from `http://www1.maths.leeds.ac.uk/~voss/2017/MATH1712/ex05-q18.csv` .)

b) In R, compute the estimates $\hat\alpha$ and $\hat\beta$ directly, using the formulas $\hat\beta = s_{xy}/s_x^2$ and $\hat\alpha = \bar y - \hat\beta \bar x$ from lectures. You can use the commands `var()` for sample variances, `cov()` for sample covariances, and `mean()` for sample means. Verify that the result you get is the same as with the `lm()` command.

c) Either by hand, or using R, produce a scatter plot showing the data together with the fitted regression line.

Note that, despite the fact that the question uses R, you still need to write your answer in complete sentences, using correct punctuation. Explain the R commands you are using and explain (and comment on) the results you get.

Exercise 19. Given $x_1, \dots, x_n, \alpha, \beta \in \mathbb{R}$, let $Y_i = \alpha + \beta x_i + \varepsilon_i$, where the ε_i are i.i.d. with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Show that the following statements hold:

a) $\mathrm{Cov}(\varepsilon_i, \bar{\varepsilon}) = \sigma^2/n$ for all $i \in \{1, \dots, n\}$, where $\bar{\varepsilon}$ is the average of the ε_i.

b) $\mathrm{Cov}(Y_i, \bar{Y}) = \sigma^2/n$ for all $i \in \{1, \dots, n\}$, where $\bar{Y}$ is the average of the Y_i.

c) $\mathrm{Cov}\big(Y_i - \bar{Y}, \bar{Y}\big) = 0$.

d) $\mathrm{Cov}(\hat{\beta}, \bar{Y}) = 0$, where $\hat{\beta}$ is the least squares estimator for β.

(Some rules for working with covariances are summarised at the end of this exercise sheet.)

Exercise 20. Assume that $X \sim B(10, p)$, *i.e.* X is binomially distributed with parameters $n = 10$ and $p \in [0, 1]$.

a) For $p = 1/2$, show that $P(X \leq 1) < 0.05$.

b) Find a value of p such that $P(X \leq 1) \geq 0.05$.

c) Let $k \in \{0, 1, \dots, n\}$. For which value of p is the probability $P(X = k)$ largest?

Reminder. If $a \in \mathbb{R}$ is a constant and X and Y are random variables, then the following rules hold for covariances.

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(a, X) = 0$.
- $\mathrm{Cov}(X, aY) = a\,\mathrm{Cov}(X, Y)$
- $\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$
- $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
- if X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$